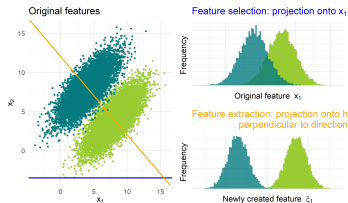


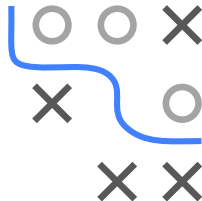
Supervised Learning

Feature Selection



Learning goals

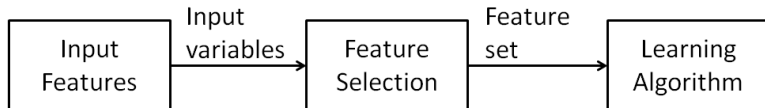
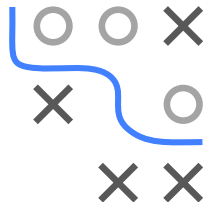
- Understand that adding more features can be detrimental to predictive performance.
- Understand the benefits of keeping only informative features for the model.



INTRODUCTION

Feature selection deals with

- techniques for choosing a suitable subset of features
- evaluating the influence of features on the model



Feature selection can be performed relying on domain knowledge and expert input, or using a data-driven algorithmic approach.

A 3x3 grid of symbols. The top row contains 'o', 'o', 'x'. The middle row contains 'x', an empty space, 'o'. The bottom row contains an empty space, 'x', 'x'. A blue line starts at the top-left corner, goes right, then down, then right, separating the 'o's from the 'x's.

-
- The graph illustrates the relationship between the number of features and the Mean Squared Error (MSE) for both training and testing data. The x-axis represents the 'Number of features' (1 to 12), and the y-axis represents the 'Mean Squared Error' (0 to 150). The green line represents the training MSE, which decreases from approximately 25 to 0 as the number of features increases. The orange line represents the test MSE, which decreases from approximately 25 to a minimum of about 10 at 10 features, then increases sharply to about 125 at 12 features, indicating overfitting.
- | Number of features | mse.train | mse.test |
|--------------------|-----------|----------|
| 1 | 25 | 25 |
| 2 | 25 | 25 |
| 3 | 15 | 15 |
| 4 | 15 | 15 |
| 5 | 15 | 15 |
| 6 | 15 | 15 |
| 7 | 15 | 15 |
| 8 | 10 | 10 |
| 9 | 5 | 20 |
| 10 | 0 | 40 |
| 11 | 0 | 40 |
| 12 | 0 | 125 |

MOTIVATION

- In high-dimensional data sets, we often have prior information that many features are either irrelevant or redundant.
- Feature selection is critical for
 - reducing noise and overfitting,
 - improving performance/generalization,
 - interpretability by identifying most informative features.
- Feature selection can also remedy problems arising in small n regimes or under limited computational resources.
- Many models require $n > p$ data. Thus, we either need to
 - adapt models to high-dimensional data (e.g. regularization),
 - design entirely new procedures for $p > n$ data, or
 - use the preprocessing methods addressed in this lecture.



SIZE OF DATASETS

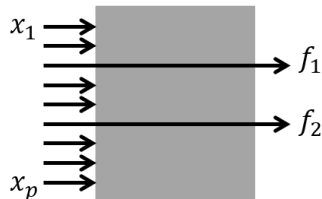
The increasingly automatized collection of information makes data sets with extremely high dimensionality available, while classical models were developed for small p data.

- **Classical setting:** Up to around 10^2 features, feature selection might be relevant, but benefits often negligible.
- **Datasets of medium to high dimensionality:** At around 10^2 to 10^3 features, classical approaches can still work well, while principled feature selection helps in many cases.
- **High-dimensional data:** 10^3 to 10^9 or more features. Examples are e.g. micro-array / gene expression data and text categorization (bag-of-words features). If, in addition, observations are few, the scenario is called $p \gg n$.



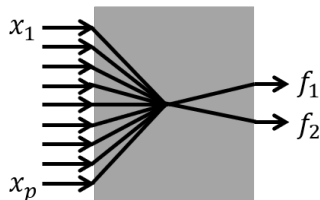
FEATURE SELECTION VS. EXTRACTION

Feature selection

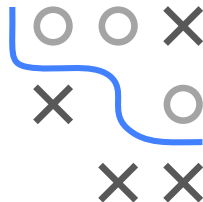


- Creates a subset of original features \mathbf{x} by selecting $\tilde{p} < p$ features \mathbf{f} .
- Retains information on selected individual features.

Feature extraction

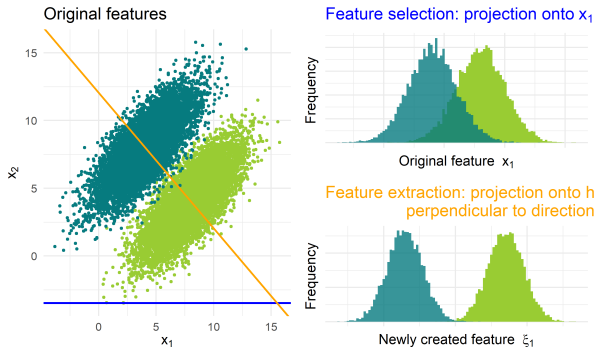
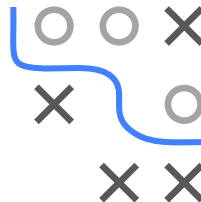


- Maps p features in \mathbf{x} to \tilde{p} extracted features \mathbf{f} .
- Info on individual features can be lost through (non-)linear combination.



FEATURE SELECTION VS. EXTRACTION

- Both FS and FE contribute to
1) dimensionality reduction, and 2) simplicity of classification rules.
- FE can be unsupervised (PCA, Multidimensional Scaling, Manifold Learning) or supervised (supervised PCA, partial least squares).
- FE can produce lower dim projections which can be more informative than FS.

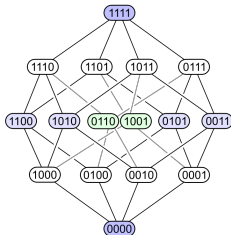
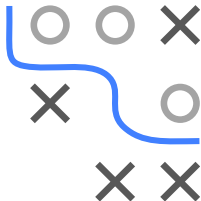


OBJECTIVE FUNCTION

Given p features, the **best-subset selection problem** is to find a subset $S \subseteq \{1, \dots, p\}$ optimizing objective $\Psi : \Omega \rightarrow \mathbb{R}$:

$$S^* \in \arg \min_{S \in \Omega} \{\Psi(S)\}$$

- Ω = search space of all feature subsets $S \subseteq \{1, \dots, p\}$. Usually we encode this by bit vectors, i.e., $\Omega = \{0, 1\}^p$ (1 = feat. selected)
- Objective Ψ can be different functions, e.g., AIC/BIC for LM or cross-validated performance of a learner.



Hasse diagram (source: Wikipedia)

HOW DIFFICULT IS BEST-SUBSET SELECTION?

- Size of search space = 2^p , i.e., grows exponentially in p as it is the power set of $\{1, \dots, p\}$.
- Finding best subset is discrete combinatorial optimization problem also known as L_0 regularization.
- It can be shown that this problem unfortunately can not be solved efficiently in general (NP-hard; see, e.g., Natarajan, 1995).
- We can avoid having to search the entire space by employing efficient search strategies, moving through the search space in a smart way that finds performant feature subsets.

